DOCUMENT RESUME

ED 202 916 TM 810 407

AUTHOR Doss, David A.
TITLE Will Removing a Few Bad Apples Save the Barrel?
PUB DATE Apr 81
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Goodness of Fit; Grade 5; Intermediate Grades; *Latent Trait Theory; Mathematical Models; *Pretests Posttests; Reading Tests; *Regression (Statistics); *Scores
IDENTIFIERS California Achievement Tests; Iowa Tests of Basic Skills; *One Parameter Model; *Rasch Model; Title I Evaluation and Reporting System

ABSTRACT

For many reasons, students do not always perform on tests in ways that are congruent with their true abilities or achievement levels. Problems of this sort may be especially common on Title I evaluations where the tests administered are too difficult for a large percentage of the students. Such considerations raise the questions of whether anything can be done to identify and remove scores which appear to be "invalid" so that some meaning can be gleaned from the evaluation. The effects of using the Rasch person-fit statistic to remove students with possibly invalid scores from a Title I Model C analysis were investigated. First a Model C analysis was computed using all students. Then those students with a reading subtest fit statistic in the top 10% district-wide were removed and the Model C analysis was redone. Then students with scores in the top 20% and the top 30% were removed. The results showed that removing some students produced a modest change in the evaluation outcome and that removing more students did not lead to significantly more change. The results support the value of using the Rasch person-fit statistic to identify students who do not fit the model. The consistent reduction of the standard error of estimate as more students were removed seems to indicate that the procedure identifies the "right" students. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED202916

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

WILL REMOVING A FEW BAD APPLES SAVE THE BARREL?

David A. Doss

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. A. Doss

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

AUSTIN INDEPENDENT SCHOOL DISTRICT
Office of Research and Evaluation

Austin, Texas

Paper presented at the Annual Meeting of the
American Educational Research Association
Los Angeles, 1981

Publication No. 80.53

TM 810 407

WILL REMOVING A FEW BAD APPLES SAVE THE BARREL?

For many reasons, students do not always perform on tests in ways that are congruent with their true abilities or achievement levels. The scores they receive may over- or underestimate their true status. One customarily assumes that the effects of such "invalid" scores on the outcome of an evaluation are minimal, random, and distributed in such a way that conclusions drawn from the data are unaffected.

Anyone long associated with the evaluation of compensatory education programs such as Title I may be excused for having some doubts about those assumptions. Measurements are often used in Title I evaluations that are made with tests that are too difficult for a large percentage of the target population. Moreover, the students may not be very strongly motivated to perform well.

Such considerations raise the question of whether anything can be done to identify and remove scores which appear to be "invalid" so that some meaning can be gleaned from the evaluation. Must the bad apples spoil the whole barrel, or can they somehow be identified and removed? And if they can be removed, can an evaluator argue that the remaining apples truly represent the whole barrel?

One approach to test theory which appears to offer hope in this direction is the Rasch one parameter, latent-trait model (Rasch, 1960). According to Wright and Stone (1979), the assumptions of the Rasch model allow the calculation of item difficulties which are independent of the sample of persons taking the test and person abilities which are independent of the items making up the test. The assumptions of the Rasch model also allow for the calculation of a measure of person fit (Wright, 1977; Mead, 1976).

To quote from Wright and Stone (1979):

> The Rasch model bases calibration and measurement on two expectations:
>
> (1) that a more able person should always have a greater probability of success on any item than a less able person, and (2) that any person should always be more likely to do better on an easier item than on a harder one. When an observed pattern of responses shows significant deviations from these expectations, we can use the particulars of the model and the person and item estimates to calculate an index of how unexpected any particular person... is. (emphasis added)

The availability of such a person-fit measure raises the possibility that the accuracy of evaluations can be improved by the removal of students with possibily invalid scores from the analyses. Does the removal of students with questionable response patterns change the conclusions one would draw about the effectiveness of a program?

To answer this question, a series of analyses was done using data from the 1979-80 evaluation of Austin's Title I Program. Specifically, the analyses examine whether removing students with possibly invalid scores affects the conclusions one would draw from a Title I Model C analysis (Tallmadge, and Wood, 1976). There were four steps to the analyses:

1. Rasch calibrate the pre- and posttests to obtain student fit statistics for each child.
2. Examine the distribution of student fit statistics to establish values for excluding students.
3. Perform the Model C analyses on all of the students. Then remove successively greater numbers of students and repeat the analyses.
4. Examine the impact of removing students on the outcome.

The paragraphs below discuss the procedure and results step-by-step.

Rasch Calibration of the Pretest and Posttest

The Model C analyses reported below were done using the Reading Total ADSS scores on the 1970 California Achievement Tests (level 3, Form A) as the pretest and the Reading Total grade equivalent scores of the 1980 Iowa Tests of Basic Skills as the posttest.

The regression analyses are limited to students in the fifth grade in Title I school in 1979-80 because that was the only grade for which item responses were available for both the pretest and posttest. The Rasch calibrations, however, were done using students who were tested district-wide.

Each of the four subtests was calibrated separately using program RASCH from the PRIME system (Veldman, 1978) at the University of Texas at Austin. The calibration of the CAT subtests used all students (grades 4 and 5) who were given Level 3 in April, 1979. Altogether 7,744 students were included. Most fifth graders in Title I schools took level 10 or 11 of the ITBS in April, 1980. These two levels were calibrated separately. The number of students used in each calibration was as follows:

- Level 10 Vocabulary, N = 3202
- Level 10 Comprehension, N = 3142
- Level 11 Vocabulary, N = 3641
- Level 11 Compehension, N = 3642

As part of the calibration, a residual mean square fit statistics or RMSQ was computed for each student for each subtest. These RMSQ statistics were used to select students to be removed from successive Model C analyses.

After calibration the data files were merged. Only students with all four subtests were used for further analyses (N = 2528).

Establish Fit Statistic Cutoffs

The greater the score a student receives on the person-fit statistic the greater the probability that the student's responses do not fit the Rasch model. According to the model, the probability of a correct response is a function of item difficulty and student ability. If the items of a test were arranged in order of increasing difficulty along a dimension, one would expect to see a pattern of correct responses followed at some point by a

pattern of incorrect responses. The point where the pattern changes from correct to incorrect would represent the student's ability or achievement level. In the real world, however, such clear patterns do not usually occur. Students tend to miss some items which they should answer correctly, and they correctly answer some items which they should miss. The RMSQ of the RASCH program quantifies the extent to which a student's response pattern deviates from the ideal (see Wright and Stone, 1979). The greater the fit statistic the greater the probability that the score is invalid.

Table 1 provides descriptive statistics for the person-fit statistics for each subtest. Note that the mean is close to 1 in each case. Also provided in Table 1 are cutoff scores which identify students scoring in the top 10%, 20%, and 30% of each distribution.

Model C Analyses

Model C of the Title I Evaluation and Reporting System (TIERS) is also called the special regression model. The models in the TIERS are used to determine the treatment effect of Title I programs. Model C is implemented as follows.

1. All students in Title I schools are given a pretest.
2. A strict cutoff score is selected on the pretest. All students at or below the cutoff score are served by the Title I program. No student above the cutoff is served.
3. All students in the school are given the posttest.
4. Two separate pre-post regression lines are calculated, one for the comparison group and one for the Title I group. The regression line for the comparison group is extended to the left of the pretest cutoff score to provide the no-treatment expectation. The distance between the two regression lines at the Title I pretest mean provides the measure of program effect.

Figure 1 shows how a plot of the regression lines would look for a successful Title I program. Note that the Title I regression line lies above the comparison group line, and both lines have the same slope.

Figure 2 shows how the analyses can look when floor or ceiling effects are encountered or when the program is differentially effective for subgroups in the Title I population. Because results like those in Figure 2 can occur, a measure of gain is made at two points: at the Title I groups pretest mean and at the pretest cutoff score. A large difference in gain as measured at the two points indicates that the regression slopes were not equal.

For this study, Model C analyses were conducted under four conditions.

1. All students in Title I schools with both pre- (CAT Reading Total ADSS scores) and posttest scores (ITBS Reading Total grade equivalent scores) were used.
2. Students with one or more subtest fit statistics in the top 10% were excluded from the Model C analyses.
3. Students with one or more fit statistics in the top 20% were excluded from the Model C analyses.
4. Students with one or more fit statistics in the top 30% were excluded from the Model C analyses.

All Model C analyses were done using program REGRAN of the PRIME system. The results for all students are presented in Table 2. The top row in the table shows that when the measurement was taken at the Title I pretest mean, the Title I equation predicted a posttest score of 24 NCE's and the comparison group equation predicted a posttest score of 27 NCE's. The NCE gain was -3 NCE's. When the regression lines were compared at the cutoff score, the difference was even greater, 9 NCE's. The two regression equations and the correlation between the CAT and ITBS for each group are found in Table 2.

It seems clear that the slopes of the two regression lines are not equal. To check that possibility, two linear models were compared using program REGRAN. The two models were as follows:

Model 1

$$Y = a_0 U + a_1 X^{(1)} + a_2 X^{(2)} + a_3 X^{(3)} + E$$

where Y is the posttest

U is the unit vector.

$X^{(1)}$ is pretest score if Title I; 0, otherwise.

$X^{(2)}$ is pretest score if non-Title I; 0, otherwise.

$X^{(3)}$ is 1 if Title I; 0, otherwise.

Model 2

$$Y = a_4U + a_5X^{(3)} + a_6X^{(4)}$$

where Y is the posttest.

U is the unit vector.

$X^{(3)}$ is 1 if Title I; 0, otherwise.

$X^{(4)}$ is pretest.

Model 1 allows the regression slopes for the two groups (Title I and non-Title I) to have separate slopes and intercepts. Model 2 requires the two groups to have a common slopes. A significant drop in the R^2 means that the slopes are not equal. Table 6 shows that the slopes were not equal.

The same analyses were then done after any students with at least one subtest fit statistic in the top 10% districtwide were removed. Then the analyses were done using the 20% and 30% cutoffs. Tables 3 through 5 report the results of these analyses.

Note how the number of students remaining in the analyses drops dramically from analysis to analysis. Table 7 shows that only 13% of the Title I students remained after those with a score in the top 30% were removed.

Does Removing Students Improve the Evaluation?

An examination of the NCE gains shows that the biggest difference comes between the first and second sets of analyses. The NCE gain measured at the mean changes from -3 to -1 when any students with fit statistics in the top 10% are removed. The gain measured at the Title I eligibility cutoff changes from -9 to -5.

When going from 10% to 20% the change is only from -1 to 0 at the pretest mean and from -5 to -4 at the Title I cutoff. There is no change from the 20% level to the 30% level.

One frequent problem with Title I evaluations is that a large percentage of the Title I group may score at or below the chance level on the pretest. In this study all students were tested on grade level with the pretest: therefore, many were given a test that was too difficult. As a rule, these lower achieving students might be expected to have higher fit statistics, because they were more apt to guess on the test. One might therefore expect the percentage of students scoring at or below chance to decrease as students are removed from the analyses.

Table 7 shows the percentages of Title I students in each analysis who were scoring at or below the chance level on the pretest in each case. The results show that the fit statistic does tend to identify large numbers of low-achievers as having invalid scores. However, higher scoring students are also selected.

The correlations between pre- and posttest were uniformly low for both Title I and non-Title I students. Yap (1980) has shown that Model C provides "practically unbiased" estimates of treatment effects; however, when pre-post correlations are low, the estimates are not efficient. Programs with true zero treatment effects might produce evaluation gains ranging from -5 to +5 NCE's. Yap also showed that sample size was related to efficiency. The larger the N, the more accurate the estimate.

Removing students with invalid scores might be expected to increase efficiency by raising the pre-post correlation. It might also be expected to decrease efficiency by lowering the N. In order to check the change in efficiency, the standard error of estimate was computed for each regression

analysis using the formula

$$SE = \sigma\sqrt{1-r^2}$$

where σ is the standard deviation of the posttest and r is the pre-post correlation. Tables 9-11 show the results for the Title I, comparison group, and combination analyses. In each set the removal of possibly invalid scores consistently improved the accuracy of the predictions even though the sample size dropped dramically as more and more students were removed. Most of the improvement would appear to come from the shrinking of the standard deviation as students were removed rather than from an increase in the correlations, because the pre-post correlations do not improve uniformly.

Conclusion

This study has investigaged the effects of using the Rasch person-fit statistics to remove students with possibly invalid scores from a Title I Model C analysis.

The results showed that removing some students produced a modest change in the evaluation outcome and that removing more students did not lead to significantly more change. Statistical manipulation is no substitute for good measurement, and creative evaluation is no substitute for a creative and productive Title I program.

The results also provide support for the validity of the Rasch person-fit statistic. The reduction in the standard error of estimates from analysis to analysis seems to indicate that the right students were being removed. At a time when test scores appear to be used more and more in making decisions about students, it would probably be useful for school districts to routinely compute the fit statistic as part of their testing programs, so students with possibly invalid scores can be identified.

References

Mead, R. J. Assessing the fit of data to the Rasch model. Paper presented at American Educational Research Association Annual Meeting, San Francisco, 1976.

Tallmadge, G. K. & Wood, C. T. User's guide: ESEA Title I evaluation and reporting system. Mountain View, California: RMC Research Corporation, 1976.

Veldman, D. The PRIME system: Computer programs for statistical analyses. Austin, Texas: Research and Development Center for Teacher Education, the University of Texas, 1978.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14.

Wright, B. D. and Stone, M. H. Best test design: Rasch measurement. Chicago: Mesa Press, 1979.

Yap, K. O. Pretest-posttest correlation and the special regression model. Paper presented at the American Psychological Association Annual Meeting, Montreal, Canada, 1980.

	Subtest	N	Mean	S.D.	Range	Cutoff Scores for the Top... 10%	20%	30%
CAT								
	Vocabulary	2528	1.047	.583	.210-6.960	1.66	1.28	1.11
	Comprehension	2528	1.024	.452	.130-5.020	1.64	1.37	1.21
ITBS								
	Vocabulary	2528	1.036	.365	.260-7.080	1.45	1.25	1.13
	Compehension	2528	1.054	.303	.470-2.930	1.42	1.25	1.14

Table 1. DESCRIPTIVE STATISTICS FOR DISTRIBUTIONS OF RMSQ FIT STATISTICS.

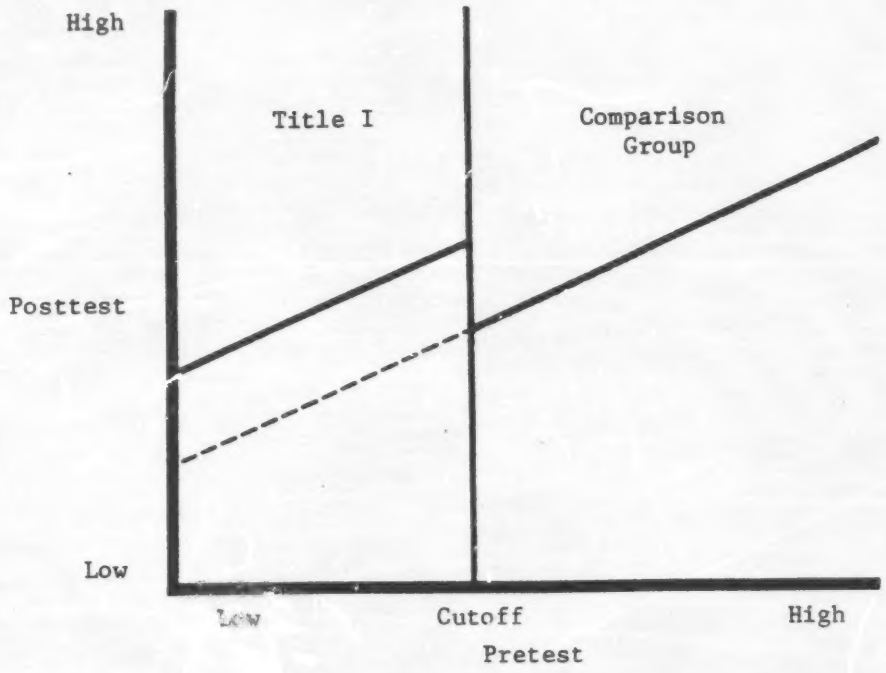

Figure 1: A SUCCESSFUL TITLE I PROGRAM.

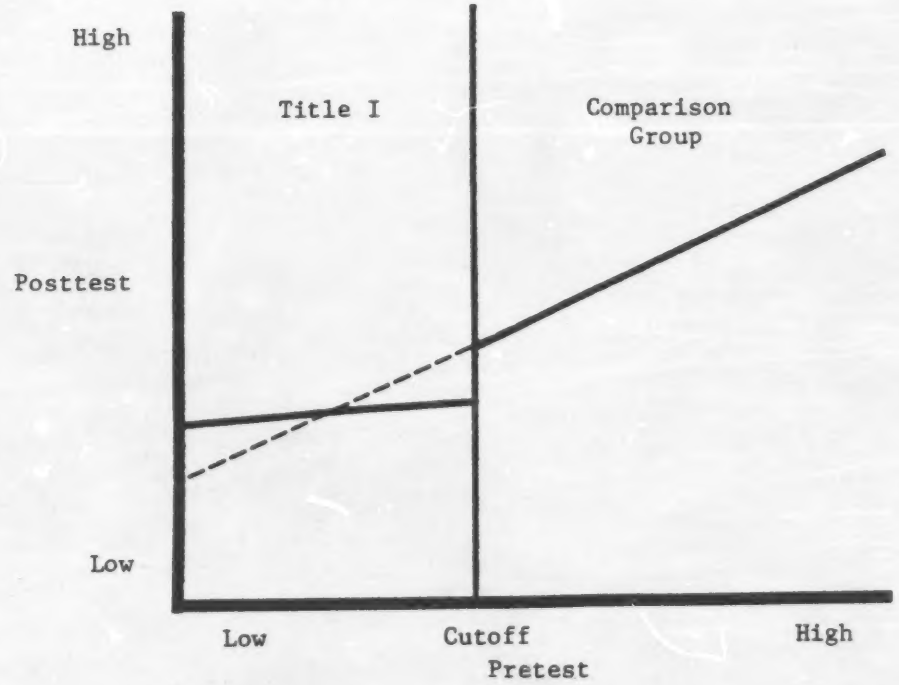

Figure 2: A MODEL C ANALYSIS WITH UNEQUAL SLOPES.

Point of Measurement	Posttest for Title I Students Predicted From... Title I Equation GE	%ile	NCE	Comparison Group Equation GE	%ile	NCE	NCE Gain
Pretest Mean (338.2)	38.5	11	24	41.3	14	27	-3
Cutoff Score (388)	45.7	20	32	51.4	33	41	-9

Comparison Group Equation: $Y = -27.3932 + .2030$ (Pretest) $r = .5842$

Title I Group Equation: $Y = -10.3501 + .1444$ (Pretest) $r = .5724$

Title I N = 441

Non-Title I N = 428

Table 2: ASSESSMENT OF TITLE I TREATMENT EFFECT USING ALL STUDENTS.

Point of Measurement	Posttest for Title I Students Predicted From... Title I Equation GE	%ile	NCE	Comparison Group Equation GE	%ile	NCE	NCE Gain
Pretest (350.8)	42.4	15	28	43.1	16	29	-1
Cutoff Score (388)	47.0	24	35	51.0	32	40	-5

Comparison Group Equation: Y = -31.6037 + .2130 (Pretest) r = .6270

Title I Group Equation: Y = -1.32 + .1245 (Pretest) r = .5039

Title I N = 198

Comparison Group N = 348

Table 3: ASSESSMENT OF TITLE I TREATMENT EFFECT USING ALL STUDENTS EXCEPT THOSE WITH ONE OR MORE STUDENT FIT STATISTICS AT OR ABOVE THE 90TH PERCENTILE OF LOCAL NORMS.

Point of Measurement	Posttest for Title I Students: Predicted from Title I Equation GE	%ile	NCE	Predicted from Non-Title I Equation GE	%ile	NCE	NCE Gain
Pretest (350.8)	44.0	18	31	43.9	18	31	0
Cutoff Score (388)	48.6	26	37	51.6	33	41	-4
Comparison Group Equation:	Y = -29.2522 + .2085 (Pretest)			r = .6234			
Title I Group Equation:	Y = .2551 + .1246 (Pretest)			r = .5126			
Title I N = 105							
Non-Title I N = 261							

Table 4: ASSESSMENT OF TITLE I TREATMENT EFFECT USING ALL STUDENTS EXCEPT THOSE WITH ONE OR MORE STUDENT FIT STATISTICS AT OR ABOVE THE 80TH PERCENTILE OF LOCAL NORMS.

Point of Measurement	Posttest for Title I Students Predicted From... Title I Equation GE	%ile	NCE	Comparison Group Equation GE	%ile	NCE	NCE Gain
Pretest (346.0)	44.3	18	31	44.3	18	31	0
Cutoff Score (388)	50.3	30	39	52.5	36	43	-4

Comparison Group Equation: Y = -23.8350 + .1968 (Pretest) r = .6317

Title I Group Equation: Y = -5.6995 + .1444 (Pretest) r = .6537

Title I N = 59

Non-Title I N = 183

Table 5: ASSESSMENT OF TITLE I TREATMENT EFFECT USING ALL STUDENTS EXCEPT THOSE WITH ONE OR MORE STUDENT FIT STATISTICS AT OR ABOVE THE 70TH PERCENTILE OF LOCAL NORMS.

Students Removed From Analyses	R^2's For Models		Test for Equal Slopes		
	Full	Restricted	df	F	P
NONE	.7071	.7030	1,865	12.04	<.001
Top 10 Percentiles	.6651	.6537	1,542	18.45	<.001
Top 20 Percentiles	.6369	.6260	1,362	10.83	.002
Top 30 Percentiles	.6698	.6650	1,238	3.42	.06*

*Test for equal intercepts (df=1,239; F=1.41; p=.23)

Table 6: TEST FOR EQUAL SLOPES.

Students Removed From the Analyses	Title I Students at or Below Chance Level	
	Number	Percent
NONE	110	24.9
Top 10 Percentiles	25	12.6
Top 20 Percentiles	14	13.3
Top 30 Percentiles	10	16.9

Table 7: NUMBER AND PERCENT OF TITLE I GROUP WHO SCORED AT OR BELOW THE CHANCE LEVEL ON THE PRETEST.

Students Removed From Analyses	Percent of Students Included		
	Title I	Non-Title I	Total
ALL	100	100	100
Top 10 Percentiles	45	81	63
Top 20 Percentiles	24	61	42
Top 30 Percentiles	13	43	28

Table 8: PERCENTAGES OF STUDENTS REMAINING IN THE ANALYSES.

Students Removed From Analyses	r	SD	SE
NONE	.5724	9.0991	7.461
Top 10 Percentiles	.5039	7.9405	6.859
Top 20 Percentiles	.5126	7.7778	6.678
Top 30 Percentiles	.6537	7.8283	5.924

r= Pre-Post Correlation
SD=Standard Deviation of Posttest
SE=Standard Error of Estimate-Posttest

Table 9: STANDARD ERRORS OF ESTIMATE FOR TITLE I GROUP EQUATIONS.

Students Removed From Analyses	r	SD	SE
NONE	.5842	9.1019	7.387
Top 10 Percentiles	.6270	7.9720	6.210
Top 20 Percentiles	.6234	7.7930	6.093
Top 30 Percentiles	.6317	7.6050	5.895

r= Pre-Post Correlation
SD=Standard Deviation of Posttest
SE=Standard Error of Estimate-Posttest

Table 10: STANDARD ERRORS OF ESTIMATE FOR COMPARISON GROUP EQUATIONS.

Students Removed From Analyses	R	SD	SE
NONE	.8409	13.7187	7.425
Top 10 Percentiles	.8155	11.1507	6.454
Top 20 Percentiles	.7981	10.4001	6.266
Top 30 Percentiles	.8104	10.2713	5.901

r= Pre-Post Correlation
SD=Standard Deviation of Posttest
SE=Standard Error of Estimate-Posttest

Table 11: STANDARD ERRORS OF ESTIMATE FOR EQUATIONS USING ALL STUDENTS.